LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

LA-UR 90-1847

1--1

LA-UR--90-1847

DE90 011998

Received
JUN

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE SAMPLING PLANS TO FIND FAULTY WASTE STORAGE DRUMS

AUTHOR(S) Richard J. Beckman and Michael D. McKay

SUBMITTED TO ASA

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy

MASTER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

FORM NO 836 R4
ST NO 2629 5/81

SAMPLING PLANS TO FIND FAULTY WASTE STORAGE DRUMS

Richard J. Beckman and
Michael D. McKay
Los Alamos National Laboratory, P.O. Box 1663,
Los Alamos, NM 87545

ABSTRACT

In environmental surveillance work, one is often faced with the task of determining by a random sample that hazardous waste is stored safely. In many cases such waste is kept in drums which are accumulated at various sites, and some of these sites are thought to have a more deleterious effect on the drums than other sites. In this paper, "knowledge" of the sites is used to develop sampling plans which increase the probability of finding faulty drums. An exchange algorithm to determine the sample sizes is given, and the influence of the prior "knowledge" on these sample sizes is investigated.

Key Words: Bayesian methods, exchange algorithm

1. Introduction

In environmental surveillance, one is often faced with the task of determining by a random sample that hazardous waste is safely stored. In many cases such waste is kept in drums that are held at various sites, and some of these sites are "thought" to have a more deleterious effect on the drums than other sites. The purpose of this paper is to develop sampling methods which take into account this "knowledge" of the sites so that we are assured that no drums are leaking.

Sampling plans which have been developed to date are concerned with the estimation of unknown quantities, such as, the total number of leaking drums (e. g. Cochran, 1977). From a regulatory point of view, this may not be the critical

criteria. In these situations, one may be more interested in guaranteeing that there are no leaking drums. Therefore, methods are developed here which maximize the probability of finding at least one leaking drum, since finding one such drum would require the storage agency to further investigate its storing practices and remedy the situation.

In the following section, a model for the generation of defective drums is developed, where the probability of a drum being defective at a site is modeled by a beta distribution. An exchange algorithm is developed to obtain the sample sizes which maximize the probability that at least one leaking drum is identified. The sensitivity of the detection probability to the prior beta distributions is investigated in Section 3. An example and a summary are given in Sections 4 and 5.

2. Model Development

Suppose there are m waste disposal sites with $N_1, N_2, ..., N_m$ storage drums. For each site, the probability that a single drum leaks is p_i, with $p_i \sim \text{Beta}(a_i, b_i)$. The values of a and b are assigned by the feeling that some sites are more prone to have leaking drums than are other sites. The model is better described as follows: It is assumed that for each site there is a probabilityp_i that a drum leaks. Given this value of p_i and the total number of drums N_i, the total leaking drums at the i^{th} site, can be viewed as a binomial with parameters p_i and N_i. The value of p_i for each site is unknown, but those doing the sampling "know" that some sites are

worse than others. This "knowledge" is not perfect, however, and this imperfect "knowledge" is reflected by allowing each p_i to have a Beta distribution.

It is probably reasonable to assume that $a_i < 1$ and $b_i > 1$ which induces an L-shaped beta distribution. Figure 1 shows a series of beta density functions for b = 25 and a = .001, .01, .25 and .5. Similar pictures may be found in Johnson and Kotz (1970). The choice of densities with small a and large b forces the density to the smaller values of p_i. While in what follows, it is not necessary to use these types of beta densities, density functions which produce large values of p_i are not as important. If all the p_i are large, then the probability of finding at least one faulty drum using any moderate sample size is almost one, no matter what the sampling plan. Therefore, on our examples and studies we will concentrate on those beta densities like those in Figure 1 where the values of p_i are almost guaranteed to be small. In those cases, the sampling strategy to optimize the probability of detection is not certain and is addressed in this paper.

We may view the process of creating a faulty drum at site i as a series of Bernoulli events, each with probability p_i. If $n_1, n_2, ..., n_m$ represent the sample sizes for the sites, with $0 \leq n_i \leq N_i$, then the probability of detecting a faulty drum, $D(\mathcal{P},\mathcal{N})$ is given by

$$D(\mathcal{P},\mathcal{N}) = 1 - \prod_{i=1}^{m}(1 - p_i)^{n_i},$$

where $\mathcal{P}^T = (p_1, p_2, ..., p_n)$ and $\mathcal{N}^T = (n_1, n_2, ..., n_m)$.

Since the vector $\mathcal{P}$ is considered a random variable, we wish to find the vector $\mathcal{N}$ which maximizes the expected value of $D(\mathcal{P},\mathcal{N})$. The expected value is

$$E_{\mathcal{P}}(D(\mathcal{P},\mathcal{N})) = 1 - E\left\{\prod_{i=1}^{m}(1 - p_i)^{n_i}\right\}$$
$$= 1 - \prod_{i=1}^{m} E\{(1 - p_i)^{n_i}\}.$$

Now,

$$E\{(1 - p_i)^{n_i}\} = \int_0^1 \left\{p_i^{a_i-1}(1 - p_i)^{b_i+n_i-1} / \beta(a_i, b_i)\right\} dp_i$$
$$= \beta(a_i, b_i + n_i)/\beta(a_i, b_i), \text{ where}$$

$$\beta(a, b) = \Gamma(a)\,\Gamma(b)\,/\Gamma(a + b).$$

Therefore,

$$E_{\mathcal{P}}(D(\mathcal{P},\mathcal{N})) = 1 - \prod_{i=1}^{m} \beta(a_i, b_i + n_i)/\beta(a_i, b_i). \qquad (1)$$

Now the vector $\mathcal{N}$ that solves $\max_{\mathcal{N}} E_{\mathcal{P}}(D(\mathcal{P},\mathcal{N}))$ is the same vector $\mathcal{N}$ that solves $\min_{\mathcal{N}} \sum_{i=1}^{m} \log \beta(a_i, b_i + n_i)$. The locally optimal vector $\mathcal{N}$ may be found by an exchange algorithm similar to those used in optimal design (e.g. Fedorov (1972)). In these algorithms a sample is deleted from the site which contributes most to $\sum_{i=1}^{m} \log \beta(a_i, b_i + n_i)$ and is added to the site which contributes least to the sum. In the first step, one finds the site that contributes most to the sum when one observation is removed from the sample. That is, the i with the maximum $\log \beta(a_i, b_i + n_i) - \log \beta(a_i, b_i + n_i - 1)$ is obtained. The $i^{\underline{th}}$ site is then a candidate to lose a sample. The second step to find the value of j which would

contribute least to the sum when an additional sample is added to that site. In other words, the j which minimizes $\log\beta(a_j, b_j + n_j + 1) - \log\beta(a_j, b_j + n_j)$ is obtained. If the loss of a sample decreases the sum more that the gain incurred by adding a sample to the j^{th} site, then n_i is decreased by one, and n_j is increased by one.

Computation of the maximum $\log\{\beta(a_i, b_i + n_i)/\beta(a_i, b_i + n_i - 1)\}$ is straight forward, since

$$\log \beta(a_i, b_i + n_i)/\beta(a_i, b_i + n_i - 1) = log\{(b_i + n_i - 1)/(a_i + b_i + n_i - 1)\}. \quad (2)$$

Now the i that maximizes (2) is the same i which minimizes $\log\{(a_i + b_i + n_i - 1)/(b_i + n_i - 1)\}$ and hence minimizes $a_i/(b_i + n_i - 1)$. In a similar manner, the j which minimizes $\log\beta(a_j, b_j + n_j + 1)/\beta(a_j, b_j + n_j)$ is the j which maximizes $a_j/(b_j + n_j)$.

These steps may be written in algorithmic form.

1. Let n_i^* (i = 1, 2,...m) be initial guesses for the sample sizes where $\sum n_i = N$, the total sample size.
2. Compute the two arrays:
   a. $DL(i) = a_i/(b_i + n_i - 1)$
   b. $DU(j) = a_j/(b_j + n_j)$.

3. For those DL(i) where $n_i>0$ find i_o where $DL(i_o)$ is a minimum.

4. For those DU(j) where $n_j<N_j$ find j_o where $DU(j_o)$ is a maximum.

5. If $DL(i_o)<DU(j_o)$ then

   a. decrease n_{i_o} by 1,

   b. increase n_{j_o} by 1,

   c. compute $DL(i_o) = a_{i_o}/(b_{i_o} + n_{i_o} - 1)$ and $DU(j_o) = a_{j_o}/(b_{j_o} + n_{j_o})$,

   d. go to step 3.

6. If $DL(i_o) \geq DU(j_o)$, the stop.

3. Sensitivity Studies

The value of a_i and b_j of (1) are chosen by the expert taking into account that some sites may be more deleterious to waste storage drums than others. Each a_i and b_j determines a prior beta distribution such as those given in Figure 1. The probability that any individual drum at site i is defective is then assumed to be p_i which is a random variable induced by the beta prior with parameters a_i and b_i. These value a_i and b_i, may be extracted from experts, Hogarth (1980), Denning (1986), and Hoffman (1987), but one would still be interested that the change in (1) if the wrong values of a_i and b_i were chosen. The purpose of this section is to investigate the sensitivity of (1) to the choice of a_i and b_i. If all of the sites are viewed as being essentially the same, then all of the a_i's and b_i's are equal. The following lemma shows that their actual values are immaterial in this case as the total sample size is split as equally as possible between the sites.

Lemma: Let J be any set of indices, i, for which the b_i's and a_j's are equal. That is for $i \in J$, then $a_i = a$ and $b_i = b$. Then for $i \in J$ the elements n_i^* of the vector $\mathcal{N}$ which maximizes (1) do not differ by more than 1.

Proof: Maximizing (1) is equivalent to minimizing

$$Q = \prod_{i=1}^{m} \Gamma(n_i + b_i)/\Gamma(n_i + a_i + b_i).$$

Without loss of generality assume that 1 and $2 \in J$ and suppose that the optimal n_1^* and n_2^* are such that $n_1^* < n_2^* + 1$ If n_1^* is reduced by 1 and n_2^* is increased by 1, Q would change by the factor

$$\Delta Q = \frac{(n_2^* + b)/(n_2^* + a)}{(n_1^* - 1 + b)/(n_1^* - 1 + a + b)}. \tag{3}$$

That is the new value of Q is $\Delta Q \times Q$.

Now both the numerator and the denominator of (3) are of the form

$$f(t) = (t + b)/(t + a + b).$$

But, since $a > 0$ and $f'(t) = a/(t + a + b)^2$, f is increasing in t. By assumption $n_1^* < n_2^* - 1$: hence ΔQ in (3) is less than 1 and Q is reduced. This contradicts the fact that n_1^* and n_2^* were optimal.

In other cases, the affect of changing the values of the a's and b's is not so clear. Therefore, a small sensitivity study was performed to judge the change in (1) if the values a_i and b_i are unspecified.

In all the studies presented here L-shaped betas were assumed. That is $a_i < 1$ and $b_i > 1$ for all i. We assumed that there were 6 sites consisting of two groups of 3, and that the total sample size would be 120. Within each group, the a_i and b_i were equal. If the incorrect values of a_i and b_i are chosen, the efficiency of the incorrect sampling plan is determined by the ratio of equation (1) with the incorrect sample allocations to equation (1) with the optimal sample allocations.

The first set of four sensitivity studies involved fixing all six a_i at the values .001, .01, .1 and .25. Figure 1 shows these densities for b = 25. It was then assumed that the true value of b was 5 for each of the six sites. Sample sizes were determined using "incorrect" values of b ranging from 2 to 100 for one of the two groups of 3 sites. The efficiencies were then determined and are given in Figure 2. As can be seen, the efficiency drops in each of these cases with the drop being the most severe for a = .001, with the lowest efficiency being .67. The flat portion of these efficiency plots is due to the fact that all of the 120 samples are allocated to the other group of sites. This is demonstrated in Figure 3 where the sample size for the first site of the first group is given as a function of b for that group. It is interesting to note that the sample size decreases approximately linearly as a function of b, and this size is independent of a. It is evident from the algorithm that this will happen as the stopping rule has DL approximately equal to DU for all sites. This basically requires that the $a_i = (b_i + n_i)$ be equal for all i. If the a_i are equal then $b_i + n_i$ must be approximately equal. Therefore,

the sample sizes are independent of a_i and the linear in b_i.

Figure 4 gives the efficiency for four fixed value of b, 2, 5, 10 and 25, and a range of values for parameters a. Once again the "true" values of a were .25 while the "true" values of b were 5. The samples were divided between two groups. The first group was assigned the (incorrectly) assumed value of a. Although the slopes of the curves in Figure 4 are interesting, from a practical point of view the efficiency does not change much as a function of a. The lowest efficiency is approximately .87. The sample sizes for the first site are given in Figure 5. As can be seen for b = 25, the sample size of the first site ranges from 0 when a is small to 40 (the maximum possible) when a approaches 1.

One may wonder what happens when beta densities such as those pictured in Figure 6 are assumed. In these cases, for any reasonable sample sizes, the expected probably of detection given by equation (1) to almost 1. Therefore, the efficiency of any sampling plan is almost 1.

4. Example

The situation consists of 9 waste storage sites that were divided into three groups. The first group, sites 1 and 2, were considered most likely to have leaking drums. The probability of a drum leaking was set at .1. We let the mean of the prior beta be a/(a + b) = .1. Choosing b = 8, requires a = .89. The next 7 sites were considered to be intermediate in probability where the mean probability was set to .01. It was then assumed for these sites that b = 10 and a = .101. The last

two sites were considered extremely unlikely to have leaking drums. The average probability was set to .001, with b = 10 and a = .01. These three densities are given in Figure 7. The optimal allocations for 250 total samples are given in the following table:

Site	N_i	a	b	n_i
1	135	.890	8	135
2	46	.890	8	46
3	1259	.101	10	14
4	176	.101	10	14
5	38	.101	10	14
6	519	.101	10	14
7	212	.101	10	13
8	349	.010	10	0
9	1056	.010	10	0

It is apparent that the bulk of the sample size is allocated to sites 1 and 2 where all of the available drums are examined. The remainder of the samples are split equally between the next 5 sites. This is in accordance with the lemma of Section 3. It is not true, however, that all L-shaped beta distributions with mean value of .1 require that the entire 135 drums of site 1 be sampled. For a/(a + b) to be equal to .1 and the resulting density be L-shaped, we only require that .11 < a < 1.0 and b = 9a. Figure 8 shows the efficiency and the sample sizes for the first site as a function of a resulting from optimal sampling plans obtained

for densities with $.11 < a < 1.0$ and $b = 9a$. For all values of a, the sample size for site 2 is always 36, while the remainder of the samples are spread equally among sites 3 to 7. One should note from Figure 8 that the sample size for the first site can be as low as 73. However, the efficiency does not drop below .98. The reason for this is the large probability of finding at least one defective, no matter what the sampling plan, as long as a substantial number of samples are taken from the first two sites.

5. Summary

In summary, we have developed a method to sample waste drums stored at different sites where we feel that some sites are more likely to have defective drums than other sites. We have shown that prior beta densities can be used to qualify our "knowledge" about the various sites with the goal to increase our chances to find at least one bad drum. An algorithm which resembles those used in optional designs gives the sample sizes to at least locally maximize the probability of detection of a faulty drum. We have also shown that the choice of the prior distribution may not greatly influence the resulting probability of detection of a faulty drum.

6. References

Denning, P. J. (1986). "The Science of Computing: Will Machines Ever Think?", *American Scientist*, Vol 74, 344–346.

Cochran, W. G. (1977). *Sampling Techniques*, New York: John Wiley and Sons.

Federov, V. V. (1972). *Theory of Optimal Experiments*, translated and edited by W. J. Studden and E. M. Klimko, New York : Academic Press.

Hoffman, R. R. (1987). "The Problem of Extracting the Knowledge of Experts from the Perspective of Experimental Psychology," *AI Magazine*, Summer, 53–66.

Hogarth, R. M. (1980). *Judgment and Choice*, New York: John Wiley and Sons.

Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions—2*, New York: John Wiley and Sons.

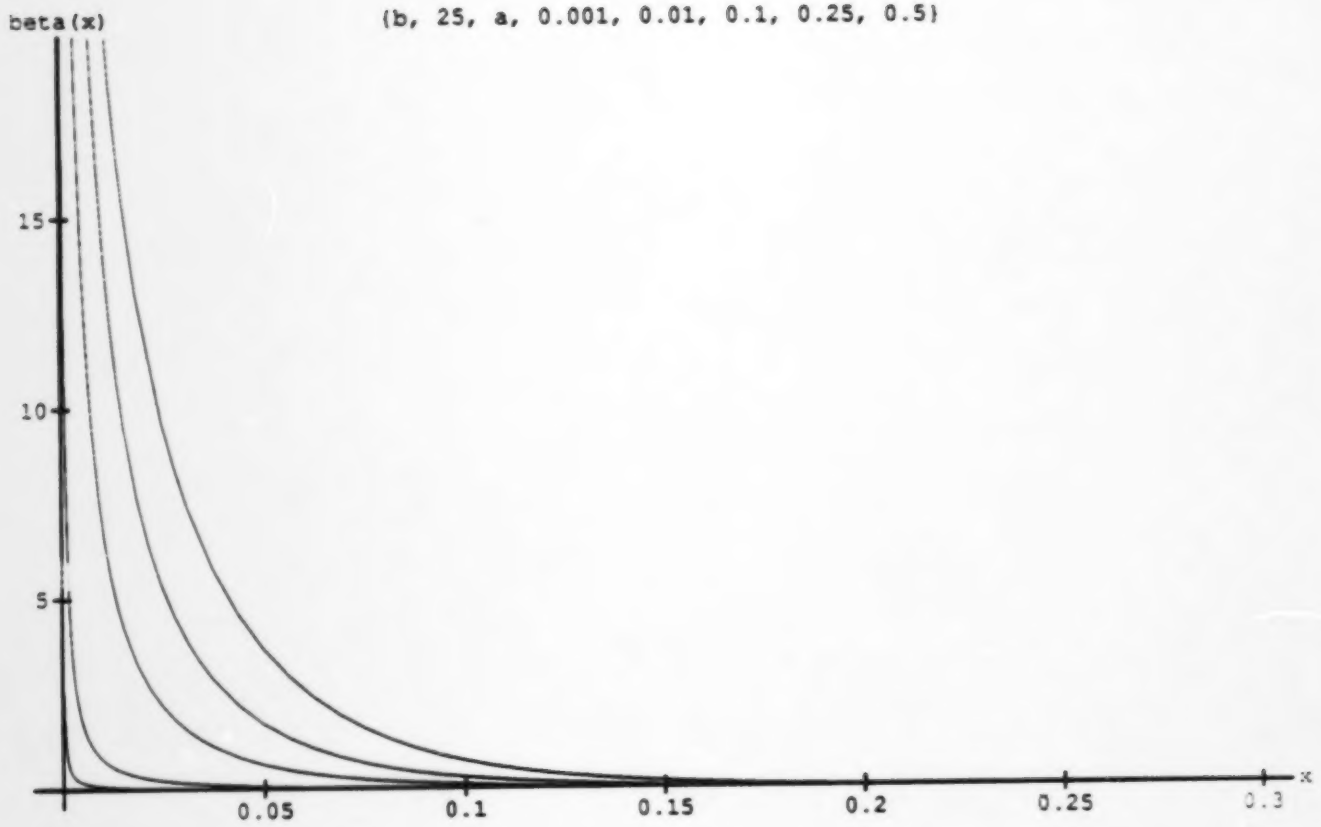


Figure 1

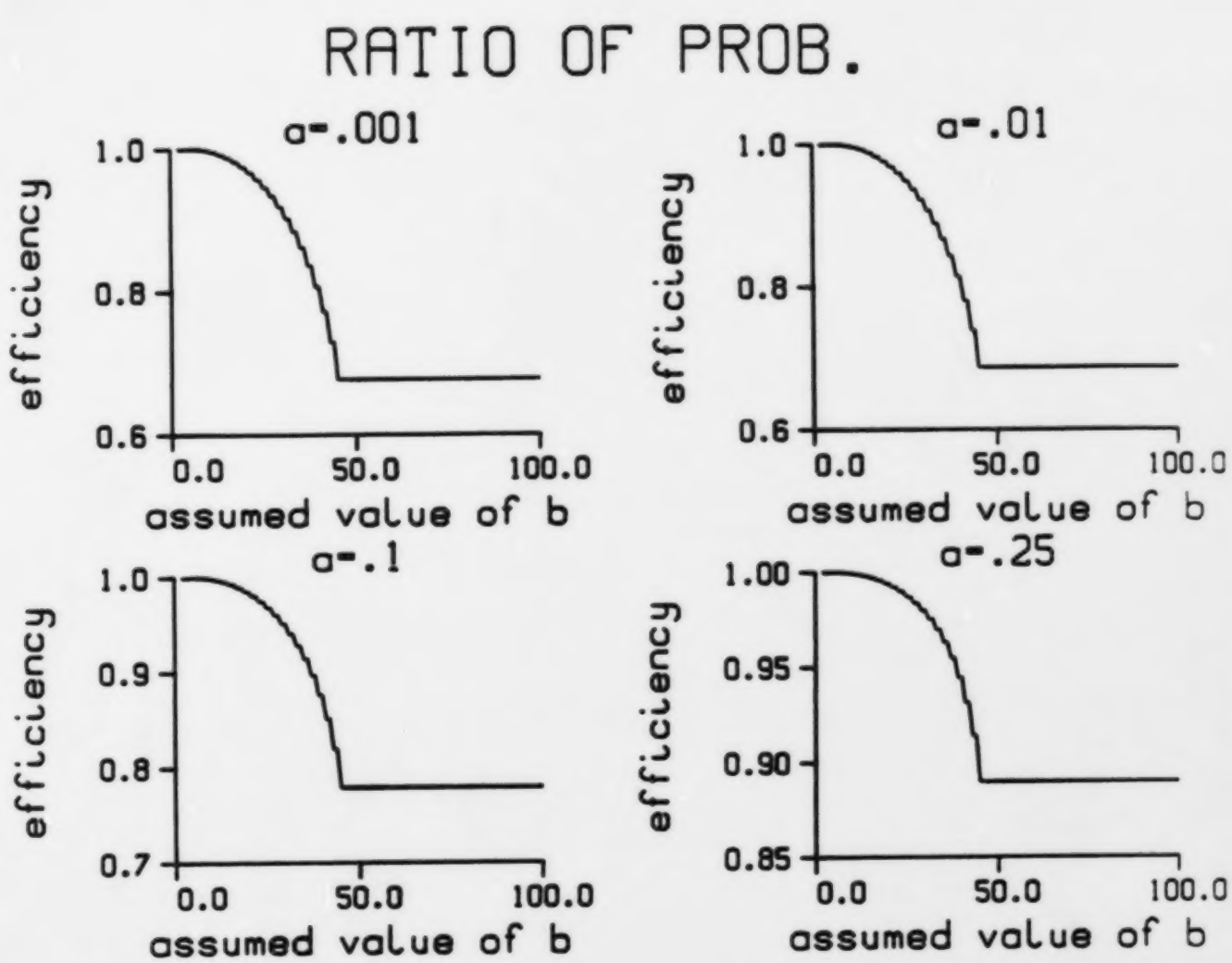

RATIO OF PROB.
a=.001
efficiency
1.0
0.8
0.6
0.0
50.0
100.0
assumed value of b
a=.01
efficiency
1.0
0.8
0.6
0.0
50.0
100.0
assumed value of b
a=.1
efficiency
1.0
0.9
0.8
0.7
0.0
50.0
100.0
assumed value of b
a=.25
efficiency
1.00
0.95
0.90
0.85
0.0
50.0
100.0
assumed value of b

Figure 2

SAMPLE SIZE FIRST SITE

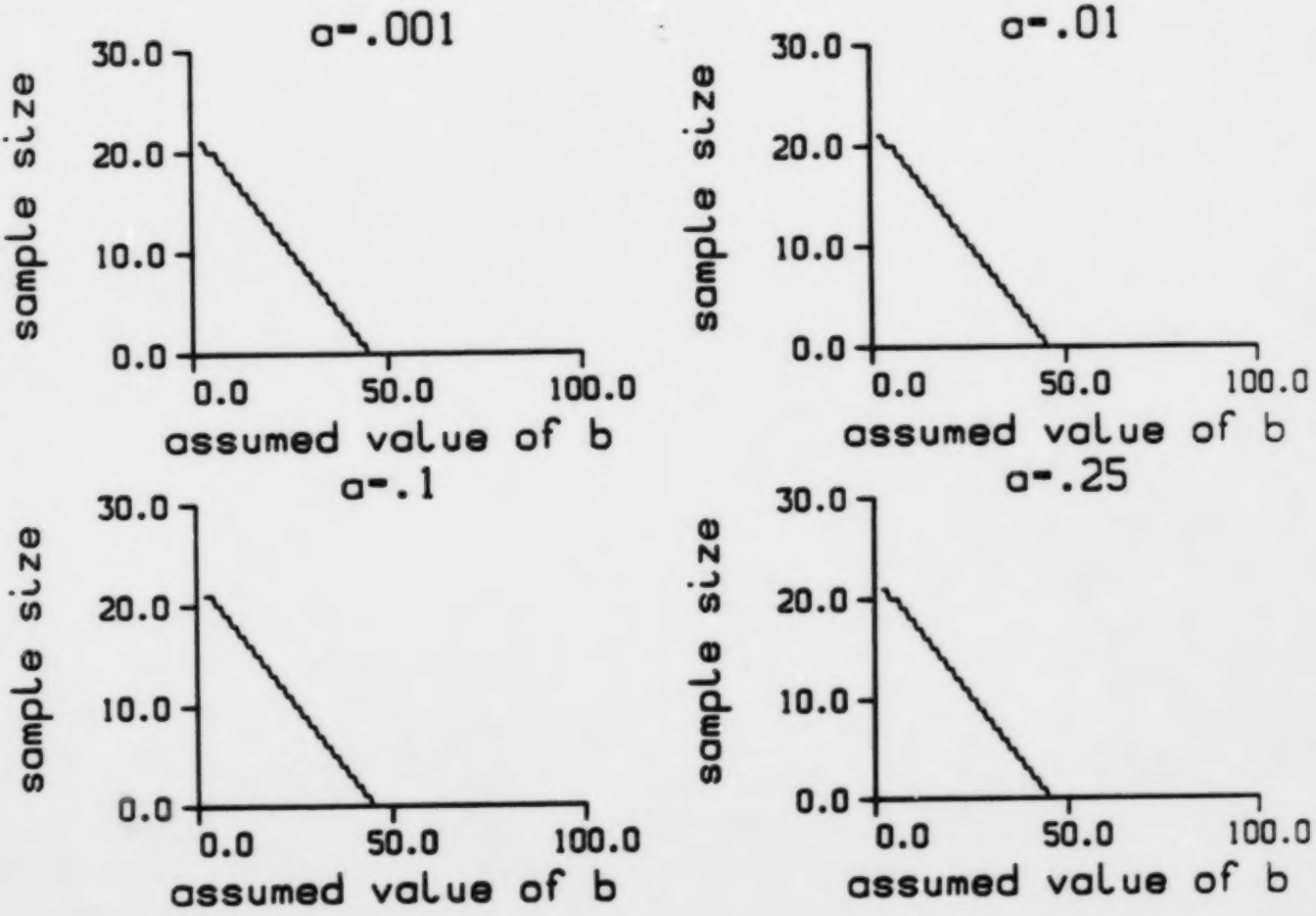


Figure 3

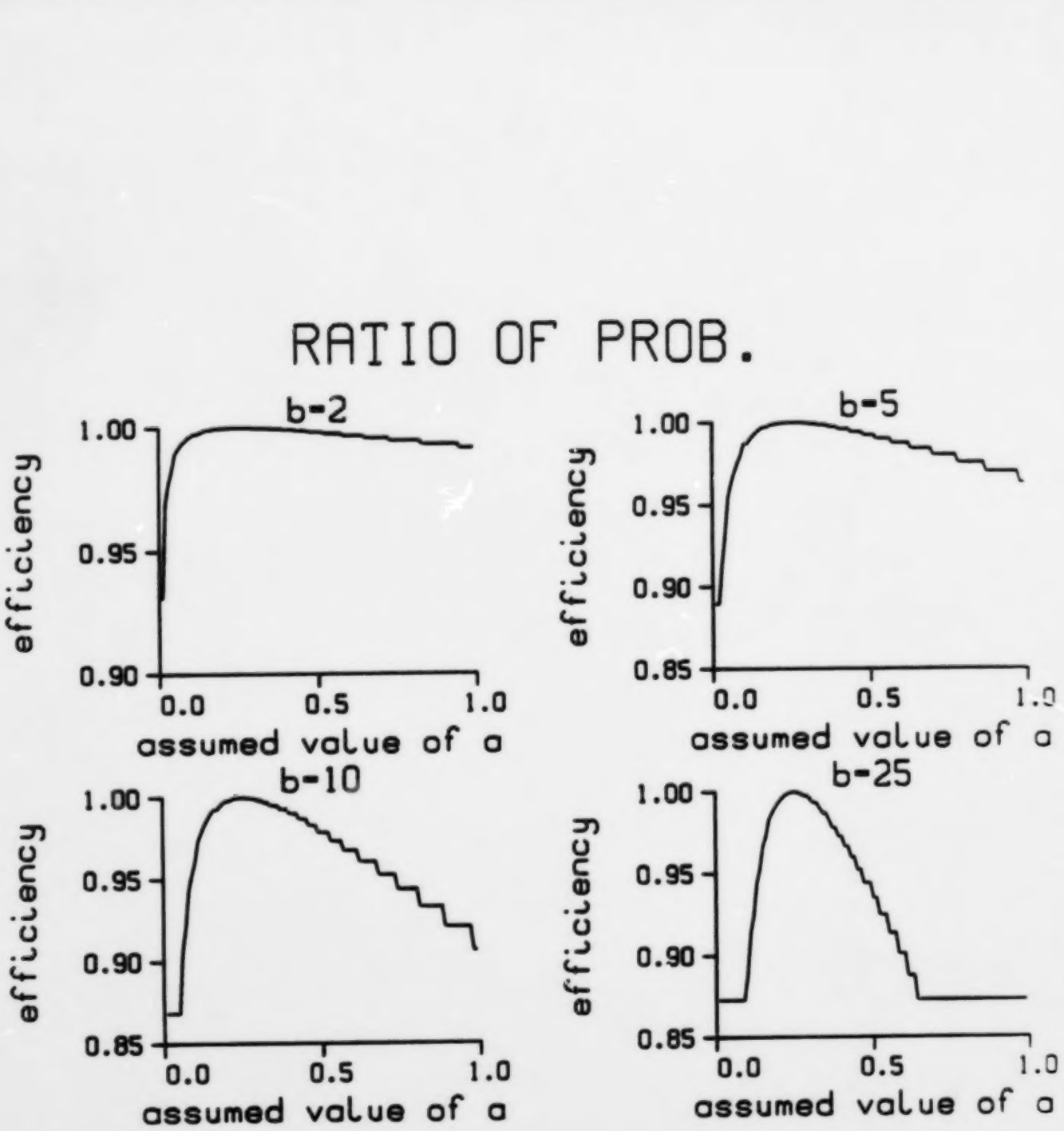
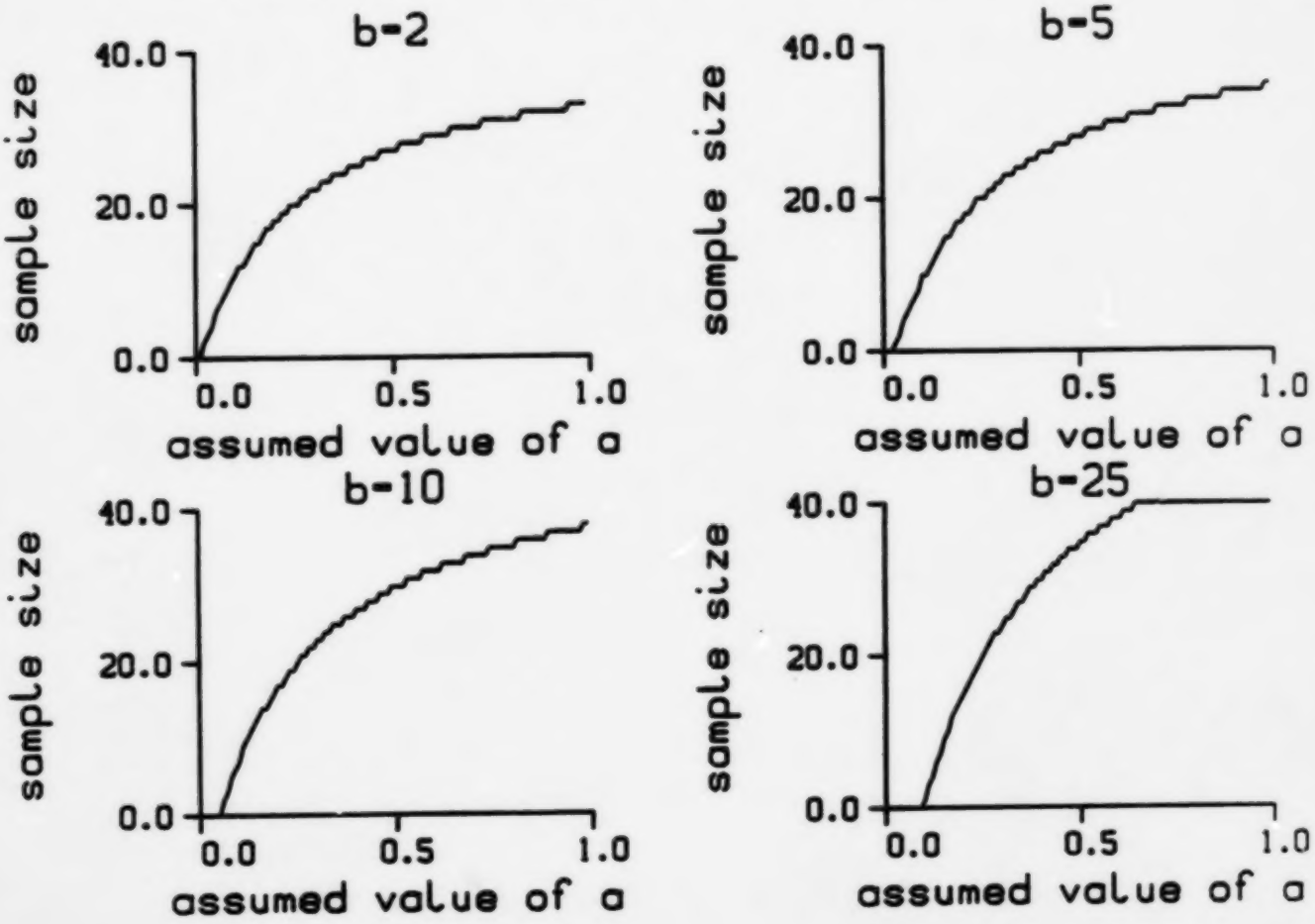

RATIO OF PROB.
b-2
b-5
b-10
b-25
efficiency
assumed value of a
1.00
0.95
0.90
0.85
0.0
0.5
1.0

Figure 4

SAMPLE SIZE FIRST SITE



Figures

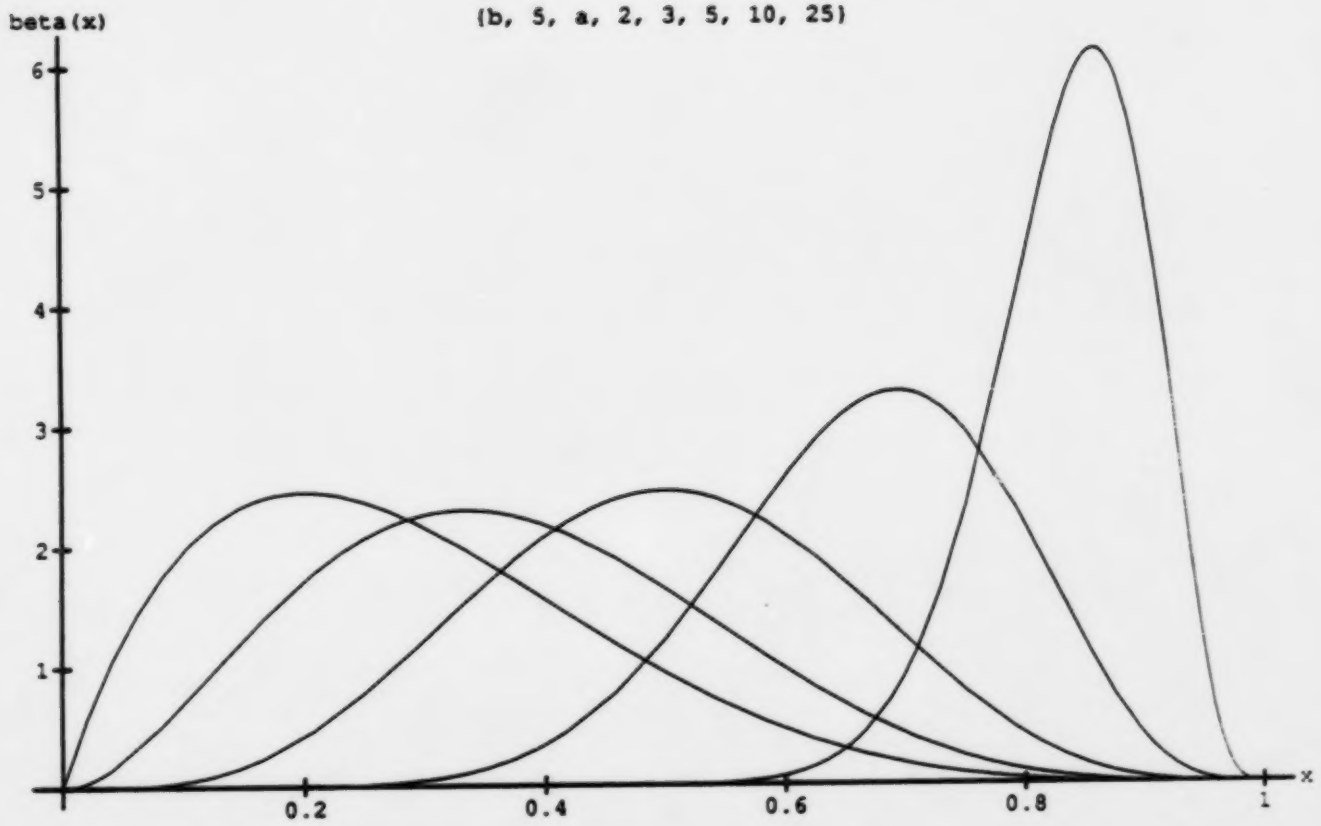


Figure 6

EXAMPLE DENSITIES

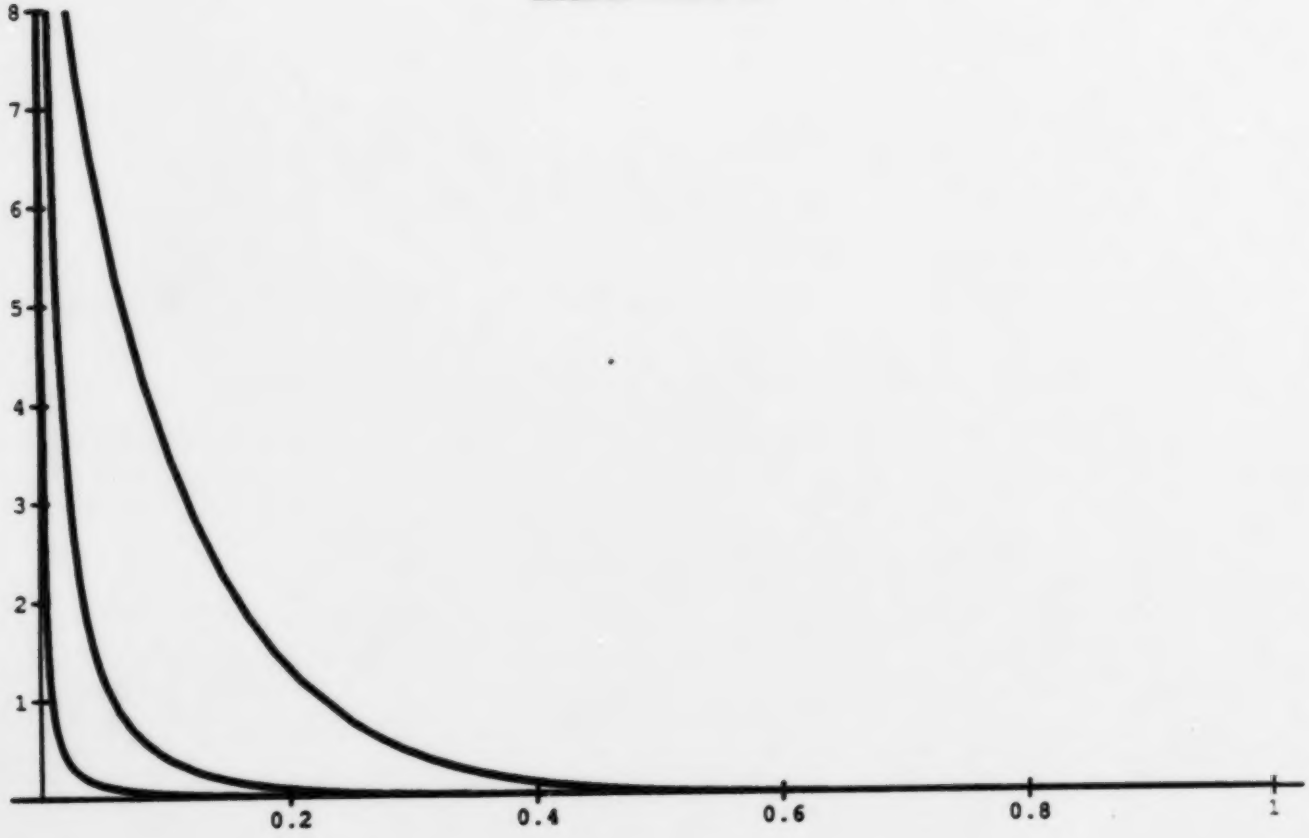


Figure 7

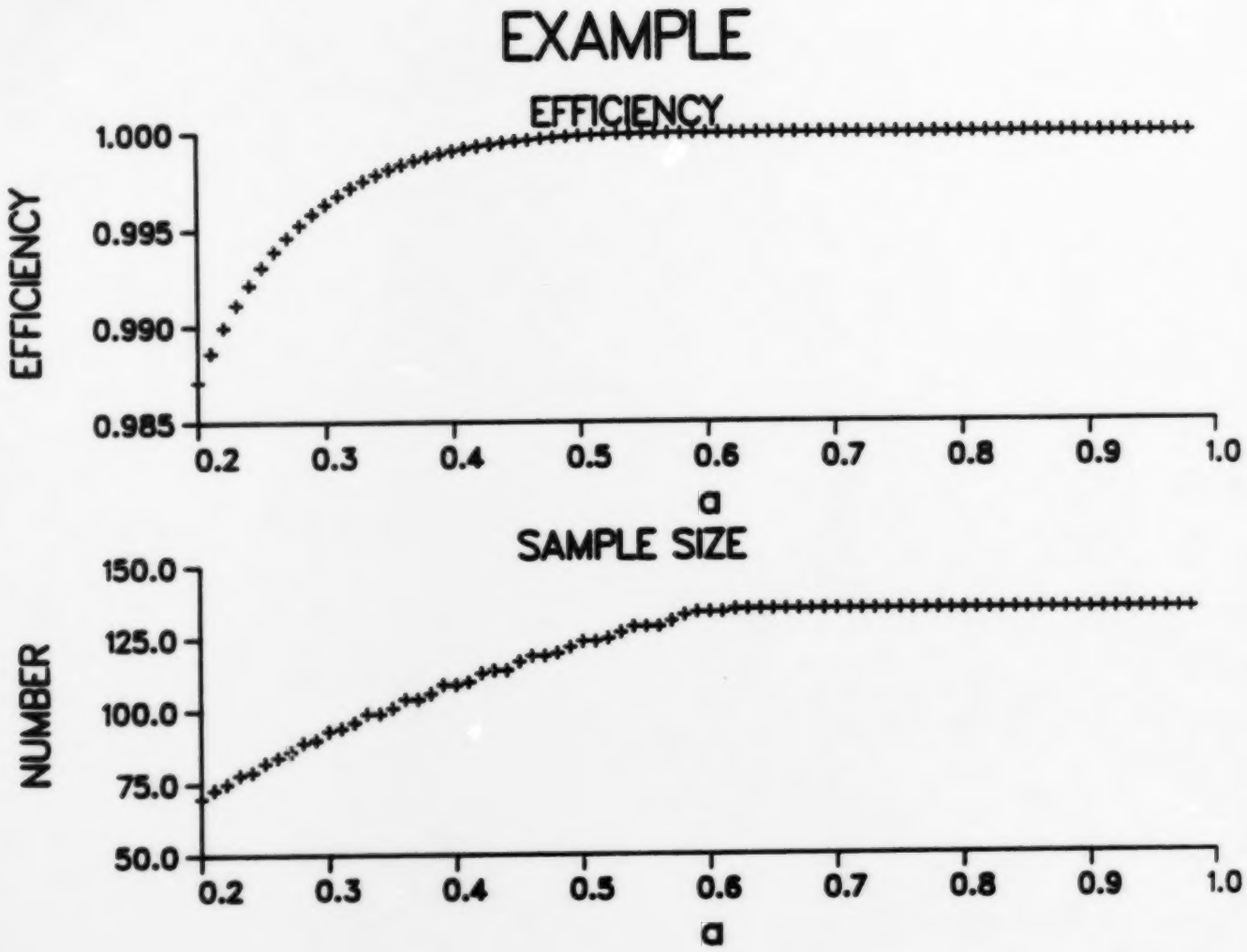

EXAMPLE
EFFICIENCY
EFFICIENCY
1.000
0.995
0.990
0.985
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
a
SAMPLE SIZE
NUMBER
150.0
125.0
100.0
75.0
50.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
a

Figure 2.